UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI
INGEGNERIA INFORMATICA,
MODELLISTICA, ELETTRONICA
E SISTEMISTICA
DIMES

Corso di Laurea Magistrale in
Ingegneria Informatica

TESI DI LAUREA

Analisi della polarizzazione politica degli utenti di Twitter mediante reti neurali

Relatori
Prof. Paolo Trunfio
Ing. Fabrizio Marozzo

Candidato
Riccardo Cantini
187900

Anno Accademico 2018/2019

Indice

Introduzione

Gli ultimi anni sono stati caratterizzati da una crescente innovazione tecnologica in moltissimi settori, nonché da un aumento esponenziale del numero dei dispositivi connessi alla rete e del loro utilizzo nella via quotidiana. Ogni utente connesso in rete lascia in maniera indelebile delle tracce a seguito della sua interazione coi vari dispositivi e le piattaforme presenti, dalle quali sono ravvisabili i suoi interessi, opinioni, abitudini di consumo o contatti. Tutto ciò porta ad un enorme insieme di dati, cui ci si riferisce col nome di *Big Data*, che vengo prodotti rapidamente e necessitano di essere acquisiti ed elaborati in tempi ridotti, trovando applicazione in diversi ambiti, dalla finanza agli e-commerce, dai trasporti alla pubblica amministrazione, dalla salute ai social media [1].

In relazione ai Big Data, negli ultimi anni sono nate una serie di tematiche d'interesse che spaziano dalla mera gestione tramite tecniche di *big data management*, supportate da appositi strumenti quali ad esempio database NoSQL come *Cassandra* o *HBase*, che garantiscono alte performance in termini di scalabilità, al *big data processing*, per il quale si annoverano vari framework open-source per il calcolo distribuito come *Apache Hadoop* o *Spark*.

In virtù delle loro caratteristiche, i Big Data sono intrinsecamente adatti ad una serie di applicazioni volte a delineare le modalità di diffusione dell'informazione, in genere all'interno di una rete, in relazione a tematiche quali *information spread* ed *influence maximization*, il comportamento abituale degli utenti, tramite ad esempio tecniche di *trajectory mining*, nonchè il loro stato d'animo o la loro opinione in riferimento ad un argomento o un evento d'interesse, tematica che afferisce alla *sentimental analysis*, in cui confluiscono diverse tecniche di *natural language processing*, in virtù della forma prevalentemente testuale in cui i dati si presentano.

In questo lavoro sono di particolare interesse i *Big Social Data* [2], ovvero il sottoinsieme dei Big Data che deriva dalle interazioni degli utenti con i social media, quali ad esempio Facebook, Twitter, Instagram, YouTube e altri, la cui diffusione è aumentata vertiginosamente, tanto da rendere il loro utilizzo parte integrante della vita quotidiana.

L'analisi dei Big Social Data, che rientra nel campo della *social media analysis*, è volta allo studio delle interazioni degli utenti sui social al fine di delinearne un profilo che ne descriva le caratteristiche salienti dal punto di vista psicologico e comportamentale: ciò consente di analizzare a fondo l'opinione pubblica, espressione della collettività, studiandola dal punto di vista della soggettività dei singoli utenti, il che permette di modellare in maniera precisa la loro percezione della società.

Il presente lavoro è incentrato sull'utilizzo dei dati social, in particolare quelli provenienti da *Twitter*, per determinare il grado di apprezzamento, presso l'opinione pubblica, di una serie di fazioni contrapposte, in vista di un evento politico d'interesse, il che equivale alla stima dei rapporti di forza tra le varie fazioni in gioco tramite il calcolo della polarizzazione dell'opinione pubblica. In particolare, viene proposta una nuova metodologia, basata su *reti neurali* (*neural networks*), volta alla stima della polarizzazione dell'opinione pubblica durante un evento politico d'interesse, la quale può essere rivolta verso una particolare fazione o un candidato, nel caso di un'elezione, o ancora verso una scelta di tipo dicotomico nel caso di un referendum.

Il presente lavoro di tesi si articola in cinque capitoli. Il primo è dedicato alla descrizione dei *social network* più diffusi e si focalizza sulle loro caratteristiche e sul loro grado di diffusione a livello mondiale. In particolare, vengono trattati i seguenti social network: *Facebook*, che detiene il primato per il maggior numero di utenti attivi; *Instagram*, che sta acquisendo una crescente popolarità presso i più giovani e relativamente al quale vengono riportate le ultime soluzioni di machine learning introdotte contro il fenomeno del cyber-bullismo; *Twitter*, utilizzato nell'ambito del presente lavoro per l'estrazione dei dati social; *YouTube*, dedicato alla condivisione di contenuti video; *LinkedIn* che permette la diffusione di contenuti specifici dedicati al mercato del lavoro. Il capitolo si conclude con una descrizione dei *Big Data*, in cui vengono delineate le loro proprietà e caratteristiche principali, nonché le modalità di analisi e le principali applicazioni, specie per i dati provenienti dai social media, oggetto del lavoro svolto nell'ambito della presente tesi.

Il secondo capitolo è incentrato sulla descrizione delle *reti neurali*, modelli afferenti al campo dell'intelligenza artificiale che derivano dai sistemi biologici le particolari caratteristiche di elaborazione delle informazioni, quali la non linearità, l'alto parallelismo, la robustezza al rumore, la tolleranza ai guasti e agli errori, l'apprendimento, nonché la loro capacità di generalizzare, e si configurano come sistemi fluidi, in grado di adattarsi ad una vasta gamma di situazioni approssimando in modo accurato funzioni altamente dimensionali, che scaturiscono da diversi ambiti applicativi, dal riconoscimento di immagini all'analisi del linguaggio naturale, dall'analisi dello stato d'animo di un individuo allo sviluppo di modelli generativi per la sintesi di dati di varia natura, quali immagini, testi o musica. All'interno del capitolo vengono descritte tre fra le tipologie di reti più usate nell'attuale stato dell'arte, ovvero il *percettrone multilivello* (MLP), le *reti convoluzionali* (CNN) ed infine un particolare tipo di reti ricorrenti, *le Long Short-Term Memory* (LSTM). Il capitolo si conclude con l'introduzione dettagliata di una serie di approcci che consentono l'applicazione delle reti neurali in ambiti quali *text mining* e *NLP*.

Nel terzo capitolo viene riportata un'analisi dettagliata dell'attuale stato dell'arte delle tecniche sviluppate nell'ambito dell'*opinion mining*. Il capitolo inizia con la descrizione delle principali tecniche di *text mining* e *sentiment analysis*, continuando con la presentazione di una serie di approcci in ambito di business, inerenti alla *customer sentiment analysis*, e politico, in cui tali tecniche hanno come obiettivo l'analisi della polarizzazione dell'opinione pubblica in relazione ad un particolare evento d'interesse, come un'elezione o un referendum. Il capitolo continua con la descrizione di due recenti lavori realizzati nel 2018, i quali propongono una metodologia di opinion mining volta alla determinazione della polarizzazione degli utenti del social network Twitter al fine di predire i risultati di eventi politici, ovvero "*A novel adaptable approach for sentiment analysis on big social data*" di El Alaoui et al., e "*Analyzing Polarization of Social Media Users and News Sites during Political Campaigns*" di Fabrizio Marozzo e Alessandro Bessi. Il capitolo si conclude con la descrizione dei principali limiti degli approcci presenti nell'attuale stato dell'arte e dei i vantaggi della tecnica proposta.

Il quarto capitolo contiene una descrizione dettagliata della metodologia proposta, basata su *reti neurali feed-forward*, la quale realizza un approccio automatico che a partire da una minima quantità di conoscenza, ovvero un sottoinsieme degli hashtag notoriamente a favore dei candidati in gioco, è in grado di espandere in maniera iterativa la propria conoscenza, fintantoché non viene raggiunto un punto massimo di saturazione delle regole apprese, ovvero si realizza una condizione di punto fisso che segna la terminazione dell'algoritmo. Il capitolo prosegue passando in rassegna le varie scelte progettuali ed implementative effettuate durante lo sviluppo dei modelli e la realizzazione dei moduli coinvolti nello sviluppo della metodologia, terminando con un esempio di applicazione sviluppato ad hoc a scopo illustrativo.

Il quinto e ultimo capitolo è adibito alla validazione della metodologia proposta tramite l'applicazione a due casi di studio reali, in particolare le *elezioni italiane del 4 marzo 2018* ed il *referendum costituzionale del 4 dicembre 2016*. I due eventi politici selezionati vengono descritti ed in seguito analizzati attraverso la metodologia di opinion mining presentata nel presente lavoro di tesi, volta alla determinazione della polarizzazione dell'opinione pubblica in riferimento a ciascun evento. Infine, vengono riportati i risultati ottenuti, confrontandoli con le percentuali reali registrate a seguito della conclusione dei rispettivi eventi oggetto d'analisi. Essi risultano estremamente vicini a quelli reali, anche più della media dei sondaggi, rivelando un'ottima accuratezza della tecnica proposta, stimata tramite l'applicazione di vari indici che misurano la bontà della predizione effettuata, quali *varianza spiegata*, *log ratio accuracy* e *coefficiente di determinazione* tendenti ad 1 ed un *errore relativo medio* trascurabile.

Bibliografia

[1] F. Fedrigo, S. Campostrini, R. Franzosi, *"Le potenzialità dell'analisi dell'utilizzo dei social network a fini di marketing. Caso studio sulle sigarette elettroniche"*, 2014

[2] E. Olshannikova, T. Olsson, J. Huhtamäki, H. Kärkkäinen, "*Conceptualizing big social data*", Journal of Big Data, 4(1), 3, 2017

[3] *https://wearesocial.com/global-digital-report-2019*

[4] *https://it.wikipedia.org/wiki/Facebook*

[5] *https://www.focus.it/tecnologia/digital-life/il-caso-cambridge-analytica-facebook-data-leak-in-7-domande-e-risposte*

[6] *https://it.wikipedia.org/wiki/Instagram*

[7] *https://www.ilfattoquotidiano.it/2018/10/10/adesso-instagram-individua-atti-di-bullismo-anche-nelle-foto/4682530/*

[8] *https://it.wikipedia.org/wiki/Twitter*

[9] *https://it.wikipedia.org/wiki/YouTube*

[10] *https://it.wikipedia.org/wiki/Diritto_d%27autore*

[11] *https://it.wikipedia.org/wiki/LinkedIn*

[12] A.K. Jain, J. Mao, "*Artificial neural networks: A tutorial*", Computer, (3), 31-44.

[13] I.A. Basheer, M. Hajmeer, "*Artificial neural network: fundamentals, computing, design, and application*", Journal of microbiological methods, 2000

[14] F. Maiani, "*Applicazioni e limiti della classificazione di immagini con reti neurali convoluzionali in dispositivi mobili*", 2017

[15] "*Understanding LSTM Networks*": http://colah.github.io/posts/2015-08-Understanding-LSTMs/, August 2017

[16] "*A Gentle Introduction to the Bag-of-Words Model*": https://machinelearningmastery.com/gentle-introduction-bag-words-model/, October 2017

[17] "*Introduction to Word Embedding and Word2Vec*": https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa, September 2018

[18] *https://neo4j.com/docs/graph-algorithms/current/algorithms/similarity-cosine/*

[19] El Alaoui et al., "*A novel adaptable approach for sentiment analysis on big social data",* Journal of Big Data, 2018

[20] T. Bhuiyan, Y. Xu, A. Jøsang, "*State-of-the-Art Review on Opinion Mining from Online Customers' Feedback*", Proceedings of the 9th Asia-Pacific Complex Systems Conference, 4-7 November 2009, Chuo University, Tokyo

[21] A. Bermingham, A. F. Smeaton, "*On Using Twitter to Monitor Political Sentiment and Predict Election Results*", Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011), 2011.

[22] F. Marozzo, A. Bessi: "*Analyzing Polarization of Social Media Users and News Sites during Political Campaigns*", 2018

[23] T. Graham et al., "*New platform, old habits? Candidates' use of Twitter during the 2010 British and Dutch general election campaigns*", New media & society 18.5 (2016): 765-783.

[24] J. M. Soler et al., "*Twitter as a Tool for Predicting Elections Results*", 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2012.

[25] P. Burnap et al., "*140 characters to victory?: Using twitter to predict the uk 2015 general election*", Electoral Studies 41 (2016): 230-233.

[26] A. Tumasjan et al., "*Predicting elections with twitter: What 140 characters reveal about political sentiment*", Fourth international AAAI conference on weblogs and social media. 2010.

[27] *https://en.wikipedia.org/wiki/Rectifier_(neural_networks)*

[28] “*A Gentle Introduction to Dropout for Regularizing Deep Neural Networks*”: https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/, December 2018

[29] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization", arXiv preprint arXiv: 1412.6980, 2014

[30] *https://it.wikipedia.org/wiki/Elezioni_politiche_italiane_del_2018*

[31] *https://it.wikipedia.org/wiki/Referendum_costituzionale_del_2016_in_Italia*

[32] *https://it.wikipedia.org/wiki/Sondaggi_ referendum_costituzionale_del_2016_in_Italia*

[33] “*Referendum costituzionale: tutti i numeri*”: https://www.youtrend.it/2016/12/09/referendum-costituzionale-tutti-numeri/, December 2016